An Unsupervised Approach to Recognizing Discourse Relations

Daniel Marcu and Abdessamad Echihabi
Information Sciences Institute and
Department of Computer Science
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA, 90292
{marcu,echihabi}@isi.edu

Abstract

We present an unsupervised approach to recognizing discourse relations of CONTRAST, EXPLANATION-EVIDENCE, CONDITION and ELABORATION that hold between arbitrary spans of texts. We show that discourse relation classifiers trained on examples that are automatically extracted from massive amounts of text can be used to distinguish between some of these relations with accuracies as high as 93%, even when the relations are not explicitly marked by cue phrases.

1 Introduction

In the field of discourse research, it is now widely agreed that sentences/clauses are usually not understood in isolation, but in relation to other sentences/clauses. Given the high level of interest in explaining the nature of these relations and in providing definitions for them (Mann and Thompson, 1988; Hobbs, 1990; Martin, 1992; Lascarides and Asher, 1993; Hovy and Maier, 1993; Knott and Sanders, 1998), it is surprising that there are no robust programs capable of identifying discourse relations that hold between arbitrary spans of text. Consider, for example, the sentence/clause pairs below.

(1)
a. Such standards would preclude arms sales to states like Libya, which is also currently subject to a U.N. embargo.

b. *But* states like Rwanda before its present crisis would still be able to legally buy arms.

(2)
a. South Africa can afford to forgo sales of guns and grenades

b. *because* it actually makes most of its profits from the sale of expensive, high-technology systems like laser-designated missiles, aircraft electronic warfare systems, tactical radios, anti-radiation bombs and battlefield mobility systems.

In these examples, the discourse markers *But* and *because* help us figure out that a CONTRAST relation holds between the text spans in (1) and an EXPLANATION-EVIDENCE relation holds between the spans in (2). Unfortunately, cue phrases do not signal all relations in a text. In the corpus of Rhetorical Structure trees (www.isi.edu/∼marcu/discourse/) built by Carlson et al. (2001), for example, we have observed that only 61 of 238 CONTRAST relations and 79 out of 307 EXPLANATION-EVIDENCE relations that hold between two adjacent clauses were marked by a cue phrase.

So what shall we do when no discourse markers are used? If we had access to robust semantic interpreters, we could, for example, infer from sentence 1.a that "cannot_buy_arms_legally(libya)", infer from sentence 1.b that "can_buy_arms_legally(rwanda)", use our background knowledge in order to infer that "similar(libya,rwanda)", and apply Hobbs's (1990) definitions of discourse relations to arrive at the conclusion that a CONTRAST relation holds between the sentences in (1). Unfortunately, the state of the art in NLP does not provide us access to semantic interpreters and general purpose knowledge bases that would support these kinds of inferences. The discourse relation definitions proposed by

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2002		**00-00-2002 to 00-00-2002**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
An Unsupervised Approach to Recognizing Discourse Relations	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Information Sciences Institute ,4676 Admiralty Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

others (Mann and Thompson, 1988; Lascarides and Asher, 1993; Knott and Sanders, 1998) are not easier to apply either because they assume the ability to automatically derive, in addition to the semantics of the text spans, the intentions and illocutions associated with them as well.

In spite of the difficulty of determining the discourse relations that hold between arbitrary text spans, it is clear that such an ability is important in many applications. First, a discourse relation recognizer would enable the development of improved discourse parsers and, consequently, of high performance single document summarizers (Marcu, 2000). In multidocument summarization (DUC, 2002), it would enable the development of summarization programs capable of identifying contradictory statements both within and across documents and of producing summaries that reflect not only the similarities between various documents, but also their differences. In question-answering, it would enable the development of systems capable of answering sophisticated, non-factoid queries, such as *"what were the causes of X?"* or *"what contradicts Y?"*, which are beyond the state of the art of current systems (TREC, 2001).

In this paper, we describe experiments aimed at building robust discourse-relation classification systems. To build such systems, we train a family of Naive Bayes classifiers on a large set of examples that are generated automatically from two corpora: a corpus of 41,147,805 English sentences that have no annotations, and BLIPP, a corpus of 1,796,386 automatically parsed English sentences (Charniak, 2000), which is available from the Linguistic Data Consortium (www.ldc.upenn.edu). We study empirically the adequacy of various features for the task of discourse relation classification and we show that some discourse relations can be correctly recognized with accuracies as high as 93%.

2 Discourse relation definitions and generation of training data

2.1 Background

In order to build a discourse relation classifier, one first needs to decide what relation definitions one is going to use. In Section 1, we simply relied on the reader's intuition when we claimed that a CONTRAST relation holds between the sentences in (1). In reality though, associating a discourse relation with a text span pair is a choice that is clearly influenced by the theoretical framework one is willing to adopt.

If we adopt, for example, Knott and Sanders's (1998) account, we would say that the relation between sentences 1.a and 1.b is ADDITIVE, because no causal connection exists between the two sentences, PRAGMATIC, because the relation pertains to illocutionary force and not to the propositional content of the sentences, and NEGATIVE, because the relation involves a CONTRAST between the two sentences. In the same framework, the relation between clauses 2.a and 2.b will be labeled as CAUSAL-SEMANTIC-POSITIVE-NONBASIC. In Lascarides and Asher's theory (1993), we would label the relation between 2.a and 2.b as EXPLANATION because the event in 2.b explains why the event in 2.a happened (perhaps by CAUSING it). In Hobbs's theory (1990), we would also label the relation between 2.a and 2.b as EXPLANATION because the event asserted by 2.b CAUSED or could CAUSE the event asserted in 2.a. And in Mann and Thompson theory (1988), we would label sentence pairs 1.a, 1.b as CONTRAST because the situations presented in them are the same in many respects (the purchase of arms), because the situations are different in some respects (Libya cannot buy arms legally while Rwanda can), and because these situations are compared with respect to these differences. By a similar line of reasoning, we would label the relation between 2.a and 2.b as EVIDENCE.

The discussion above illustrates two points. First, it is clear that although current discourse theories are built on fundamentally different principles, they all share some common intuitions. Sure, some theories talk about "negative polarity" while others about "contrast". Some theories refer to "causes", some to "potential causes", and some to "explanations". But ultimately, all these theories acknowledge that there are such things as CONTRAST, CAUSE, and EXPLANATION relations. Second, given the complexity of the definitions these theories propose, it is clear why it is difficult to build programs that recognize such relations in unrestricted texts. Current NLP techniques do not enable us to reliably infer from sen-

tence 1.a that "cannot_buy_arms_legally(libya)" and do not give us access to general purpose knowledge bases that assert that "similar(libya,rwanda)".

The approach we advocate in this paper is in some respects less ambitious than current approaches to discourse relations because it relies upon a much smaller set of relations than those used by Mann and Thompson (1988) or Martin (1992). In our work, we decide to focus only on four types of relations, which we call: CONTRAST, CAUSE-EXPLANATION-EVIDENCE (CEV), CONDITION, and ELABORATION. (We define these relations in Section 2.2.) In other respects though, our approach is more ambitious because it focuses on the problem of recognizing such discourse relations in unrestricted texts. In other words, given as input sentence pairs such as those shown in (1)–(2), we develop techniques and programs that label the relations that hold between these sentence pairs as CONTRAST, CAUSE-EXPLANATION-EVIDENCE, CONDITION, ELABORATION or NONE-OF-THE-ABOVE, *even when the discourse relations are not explicitly signalled by discourse markers.*

2.2 Discourse relation definitions

The discourse relations we focus on are defined at a much coarser level of granularity than in most discourse theories. For example, we consider that a CONTRAST relation holds between two text spans if one of the following relations holds: CONTRAST, ANTITHESIS, CONCESSION, or OTHERWISE, as defined by Mann and Thompson (1988), CONTRAST or VIOLATED EXPECTATION, as defined by Hobbs (1990), or any of the relations characterized by this regular expression of cognitive primitives, as defined by Knott and Sanders (1998): (CAUSAL | ADDITIVE) – (SEMANTIC | PRAGMATIC) – NEGATIVE. In other words, in our approach, we do not distinguish between contrasts of semantic and pragmatic nature, contrasts specific to violated expectations, etc. Table 1 shows the definitions of the relations we considered.

The advantage of operating with coarsely defined discourse relations is that it enables us to automatically construct relatively low-noise datasets that can be used for learning. For example, by extracting sentence pairs that have the keyword "But" at the beginning of the second sentence, as the sentence pair shown in (1), we can automatically collect many examples of CONTRAST relations. And by extracting sentences that contain the keyword "because", we can automatically collect many examples of CAUSE-EXPLANATION-EVIDENCE relations. As previous research in linguistics (Halliday and Hasan, 1976; Schiffrin, 1987) and computational linguistics (Marcu, 2000) show, some occurrences of "but" and "because" do not have a discourse function; and others signal other relations than CONTRAST and CAUSE-EXPLANATION. So we can expect the examples we extract to be noisy. However, empirical work of Marcu (2000) and Carlson et al. (2001) suggests that the majority of occurrences of "but", for example, do signal CONTRAST relations. (In the RST corpus built by Carlson et al. (2001), 89 out of the 106 occurrences of "but" that occur at the beginning of a sentence signal a CONTRAST relation that holds between the sentence that contains the word "but" and the sentence that precedes it.) Our hope is that simple extraction methods are sufficient for collecting low-noise training corpora.

2.3 Generation of training data

In order to collect training cases, we mined in an unsupervised manner two corpora. The first corpus, which we call *Raw*, is a corpus of 1 billion words of unannotated English (41,147,805 sentences) that we created by catenating various corpora made available over the years by the Linguistic Data Consortium. The second, called *BLIPP*, is a corpus of only 1,796,386 sentences that were parsed automatically by Charniak (2000). We extracted from both corpora all adjacent sentence pairs that contained the cue phrase "But" at the beginning of the second sentence and we automatically labeled the relation between the two sentence pairs as CONTRAST. We also extracted all the sentences that contained the word "but" in the middle of a sentence; we split each extracted sentence into two spans, one containing the words from the beginning of the sentence to the occurrence of the keyword "but" and one containing the words from the occurrence of "but" to the end of the sentence; and we labeled the relation between the two resulting text spans as CONTRAST as well.

Table 2 lists some of the cue phrases we used in order to extract CONTRAST, CAUSE-EXPLANATION-EVIDENCE, ELABORATION, and

CONTRAST	CAUSE-EXPLANATION-EVIDENCE	ELABORATION	CONDITION
ANTITHESIS (M&T) CONCESSION (M&T) OTHERWISE (M&T) CONTRAST (M&T) VIOLATED EXPECTATION (Ho) (CAUSAL | ADDITIVE) - (SEMANTIC | PRAGMATIC) - NEGATIVE (K&S)	EVIDENCE (M&T) VOLITIONAL-CAUSE (M&T) NONVOLITIONAL-CAUSE (M&T) VOLITIONAL-RESULT (M&T) NONVOLITIONAL-RESULT (M&T) EXPLANATION (Ho) RESULT (A&L) EXPLANATION (A&L) CAUSAL - (SEMANTIC | PRAGMATIC) - POSITIVE (K&S)	ELABORATION (M&T) EXPANSION (Ho) EXEMPLIFICATION (Ho) ELABORATION (A&L)	CONDITION (M&T)

Table 1: Relation definitions as union of definitions proposed by other researchers (M&T – (Mann and Thompson, 1988); Ho – (Hobbs, 1990); A&L – (Lascarides and Asher, 1993); K&S – (Knott and Sanders, 1998)).

CONTRAST – 3,881,588 examples [BOS . . . EOS] [BOS But . . . EOS] [BOS . . .] [but . . . EOS] [BOS . . .] [although . . . EOS] [BOS Although . . . ,] [. . . EOS]
CAUSE-EXPLANATION-EVIDENCE — 889,946 examples [BOS . . .] [because . . . EOS] [BOS Because . . . ,] [. . . EOS] [BOS . . . EOS] [BOS Thus, . . . EOS]
CONDITION — 1,203,813 examples [BOS If . . . ,] [. . . EOS] [BOS If . . .] [then . . . EOS] [BOS . . .] [if . . . EOS]
ELABORATION — 1,836,227 examples [BOS . . . EOS] [BOS . . . for example . . . EOS] [BOS . . .] [which . . . ,]
NO-RELATION-SAME-TEXT — 1,000,000 examples Randomly extract two sentences that are more than 3 sentences apart in a given text.
NO-RELATION-DIFFERENT-TEXTS — 1,000,000 examples Randomly extract two sentences from two different documents.

Table 2: Patterns used to automatically construct a corpus of text span pairs labeled with discourse relations.

CONDITION relations and the number of examples extracted from the Raw corpus for each type of discourse relation. In the patterns in Table 2, the symbols BOS and EOS denote BeginningOfSentence and EndOfSentence boundaries, the ". . . " stand for occurrences of any words and punctuation marks, the square brackets stand for text span boundaries, and the other words and punctuation marks stand for the cue phrases that we used in order to extract discourse relation examples. For example, the pattern [BOS Although . . . ,] [. . . EOS] is used in order to extract examples of CONTRAST relations that hold between a span of text delimited to the left by the cue phrase "Although" occurring in the beginning of a sentence and to the right by the first occurrence of a comma, and a span of text that contains the rest of the sentence to which "Although" belongs.

We also extracted automatically 1,000,000 examples of what we hypothesize to be non-relations, by randomly selecting non-adjacent sentence pairs that are at least 3 sentences apart in a given text. We label such examples NO-RELATION-SAME-TEXT. And we extracted automatically 1,000,000 examples of what we hypothesize to be cross-document non-relations, by randomly selecting two sentences from distinct documents. As in the case of CONTRAST and CONDITION, the NO-RELATION examples are also noisy because long distance relations are common in well-written texts.

3 Determining discourse relations using Naive Bayes classifiers

We hypothesize that we can determine that a CONTRAST relation holds between the sentences in (3) even if we cannot semantically interpret the two sentences, simply because our background knowledge tells us that *good* and *fails* are good indicators of contrastive statements.

- John is *good* in math and sciences. (3)
- Paul *fails* almost every class he takes.

Similarly, we hypothesize that we can determine that a CONTRAST relation holds between the sentences

in (1), because our background knowledge tells us that *embargo* and *legally* are likely to occur in contexts of opposite polarity. In general, we hypothesize that lexical item pairs can provide clues about the discourse relations that hold between the text spans in which the lexical items occur.

To test this hypothesis, we need to solve two problems. First, we need a means to acquire vast amounts of background knowledge from which we can derive, for example, that the word pairs *good – fails* and *embargo – legally* are good indicators of CONTRAST relations. The extraction patterns described in Table 2 enable us to solve this problem.[1] Second, given vast amounts of training material, we need a means to learn which pairs of lexical items are likely to co-occur in conjunction with each discourse relation and a means to apply the learned parameters to any pair of text spans in order to determine the discourse relation that holds between them. We solve the second problem in a Bayesian probabilistic framework.

We assume that a discourse relation r_k that holds between two text spans, W_1, W_2, is determined by the word pairs in the cartesian product defined over the words in the two text spans $(w_i, w_j) \in W_1 \times W_2$. In general, a word pair $(w_i, w_j) \in W_1 \times W_2$ can "signal" any relation r_k. We determine the most likely discourse relation that holds between two text spans W_1 and W_2 by taking the maximum over $argmax_{r_k} P(r_k|W_1, W_2)$, which according to Bayes rule, amounts to taking the maximum over $argmax_{r_k}[logP(W_1, W_2|r_k) + logP(r_k)]$. If we assume that the word pairs in the cartesian product are independent, $P(W_1, W_2|r_k)$ is equivalent to $\prod_{(w_i,w_j)\in W_1,W_2} P((w_i, w_j) \mid r_k)$. The values $P((w_i, w_j) \mid r_k)$ are computed using maximum likelihood estimators, which are smoothed using the Laplace method (Manning and Schütze, 1999).

For each discourse relation pair r_a, r_b, we train a word-pair-based classifier using the automatically derived training examples in the Raw corpus, from which we *first removed the cue-phrases used for extracting the examples.* This ensures that our classifiers do not learn, for example, that the word pair *if – then* is a good indicator of a CONDITION relation, which would simply amount to learning to distinguish between the extraction patterns used to construct the corpus. We test each classifier on a test corpus of 5000 examples labeled with r_a and 5000 examples labeled with r_b, which ensures that the baseline is the same for all combinations r_a and r_b, namely 50%.

Table 3 shows the performance of all discourse relation classifiers. As one can see, each classifier outperforms the 50% baseline, with some classifiers being as accurate as that that distinguishes between CAUSE-EXPLANATION-EVIDENCE and ELABORATION relations, which has an accuracy of 93%. We have also built a six-way classifier to distinguish between all six relation types. This classifier has a performance of 49.7%, with a baseline of 16.67%, which is achieved by labeling all relations as CONTRASTS.

We also examined the learning curves of various classifiers and noticed that, for some of them, the addition of training examples does not appear to have a significant impact on their performance. For example, the classifier that distinguishes between CONTRAST and CAUSE-EXPLANATION-EVIDENCE relations has an accuracy of 87.1% when trained on 2,000,000 examples and an accuracy of 87.3% when trained on 4,771,534 examples. We hypothesized that the flattening of the learning curve is explained by the noise in our training data and the vast amount of word pairs that are not likely to be good predictors of discourse relations.

To test this hypothesis, we decided to carry out a second experiment that used as predictors only a subset of the word pairs in the cartesian product defined over the words in two given text spans. To achieve this, we used the patterns in Table 2 to extract examples of discourse relations from the BLIPP corpus. As expected, the BLIPP corpus yielded much fewer learning cases: 185,846 CONTRAST; 44,776 CAUSE-EXPLANATION-EVIDENCE; 55,699 CONDITION; and 33,369 ELABORATION relations. To these examples, we added 58,000 NO-RELATION-SAME-TEXT and 58,000 NO-RELATION-DIFFERENT-TEXTS relations.

To each text span in the BLIPP corpus corresponds a parse tree (Charniak, 2000). We wrote

[1] Note that relying on the list of antonyms provided by Wordnet (Fellbaum, 1998) is not enough because the semantic relations in Wordnet are not defined across word class boundaries. For example, Wordnet does not list the "antonymy"-like relation between *embargo* and *legally*.

	CONTRAST	CEV	COND	ELAB	NO-REL-SAME-TEXT	NO-REL-DIFF-TEXTS
CONTRAST	-	87	74	82	64	64
CEV			76	93	75	74
COND				89	69	71
ELAB					76	75
NO-REL-SAME-TEXT						64

Table 3: Performances of classifiers trained on the Raw corpus. The baseline in all cases is 50%.

	CONTRAST	CEV	COND	ELAB	NO-REL-SAME-TEXT	NO-REL-DIFF-TEXTS
CONTRAST	-	62	58	78	64	72
CEV			69	82	64	68
COND				78	63	65
ELAB					78	78
NO-REL-SAME-TEXT						66

Table 4: Performances of classifiers trained on the BLIPP corpus. The baseline in all cases is 50%.

a simple program that extracted the nouns, verbs, and cue phrases in each sentence/clause. We call these the *most representative words* of a sentence/discourse unit. For example, the most representative words of the sentence in example (4), are those shown in italics.

Italy's unadjusted industrial *production fell* in *January* 3.4% from a *year* earlier *but rose* 0.4% from *December*, the *government said* (4)

We repeated the experiment we carried out in conjunction with the Raw corpus on the data derived from the BLIPP corpus as well. Table 4 summarizes the results.

Overall, the performance of the systems trained on the most representative word pairs in the BLIPP corpus is clearly lower than the performance of the systems trained on all the word pairs in the Raw corpus. But a direct comparison between two classifiers trained on different corpora is not fair because with just 100,000 examples per relation, the systems trained on the Raw corpus are much worse than those trained on the BLIPP data. The learning curves in Figure 1 are illuminating as they show that if one uses as features only the most representative word pairs, one needs only about 100,000 training examples to achieve the same level of performance one achieves using 1,000,000 training examples and features defined over all word pairs. Also, since the learning curve for the BLIPP corpus is steeper than

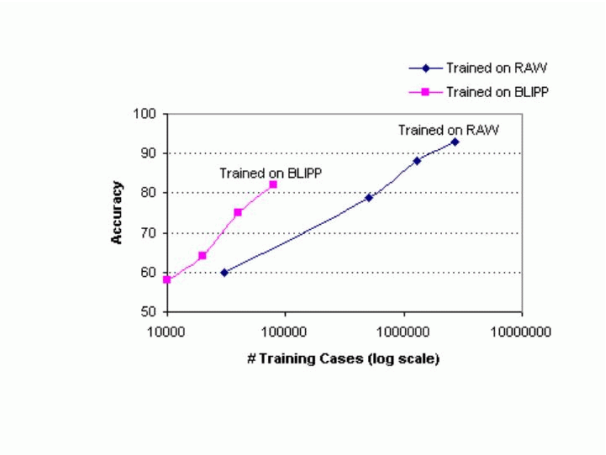

Figure 1: Learning curves for the ELABORATION vs. CAUSE-EXPLANATION-EVIDENCE classifiers, trained on the Raw and BLIPP corpora.

the learning curve for the Raw corpus, this suggests that discourse relation classifiers trained on most representative word pairs and millions of training examples can achieve higher levels of performance than classifiers trained on all word pairs (unannotated data).

4 Relevance to RST

The results in Section 3 indicate clearly that massive amounts of automatically generated data can be used to distinguish between discourse relations defined as discussed in Section 2.2. What the experiments

	CONTR	CEV	COND	ELAB
# test cases	238	307	125	1761
CONTR	—	**63** 56	**80** 65	**64** 88
CEV			**87** 71	**76** 85
COND				**87** 93

Table 5: Performances of Raw-trained classifiers on manually labeled RST relations that hold between elementary discourse units. Performance results are shown in bold; baselines are shown in normal fonts.

in Section 3 do not show is whether the classifiers built in this manner can be of any use in conjunction with some established discourse theory. To test this, we used the corpus of discourse trees built in the style of RST by Carlson et al. (2001). We automatically extracted from this manually annotated corpus all CONTRAST, CAUSE-EXPLANATION-EVIDENCE, CONDITION and ELABORATION relations that hold between two adjacent elementary discourse units. Since RST (Mann and Thompson, 1988) employs a finer grained taxonomy of relations than we used, we applied the definitions shown in Table 1. That is, we considered that a CONTRAST relation held between two text spans if a human annotator labeled the relation between those spans as ANTITHESIS, CONCESSION, OTHERWISE or CONTRAST. We retrained then all classifiers on the Raw corpus, but this time without removing from the corpus the cue phrases that were used to generate the training examples. We did this because when trying to determine whether a CONTRAST relation holds between two spans of texts separated by the cue phrase "but", for example, we want to take advantage of the cue phrase occurrence as well. We employed our classifiers on the manually labeled examples extracted from Carlson et al.'s corpus (2001). Table 5 displays the performance of our two way classifiers for relations defined over elementary discourse units. The table displays in the second row, for each discourse relation, the number of examples extracted from the RST corpus. For each binary classifier, the table lists in bold the accuracy of our classifier and in non-bold font the majority baseline associated with it.

The results in Table 5 show that the classifiers learned from automatically generated training data can be used to distinguish between certain types of RST relations. For example, the results show that the classifiers can be used to distinguish between CONTRAST and CAUSE-EXPLANATION-EVIDENCE relations, as defined in RST, but not so well between ELABORATION and any other relation. This result is consistent with the discourse model proposed by Knott et al. (2001), who suggest that ELABORATION relations are too ill-defined to be part of any discourse theory.

The analysis above is informative only from a machine learning perspective. From a linguistic perspective though, this analysis is not very useful. If no cue phrases are used to signal the relation between two elementary discourse units, an automatic discourse labeler can at best guess that an ELABORATION relation holds between the units, because ELABORATION relations are the most frequently used relations (Carlson et al., 2001). Fortunately, with the classifiers described here, one can label some of the unmarked discourse relations correctly.

For example, the RST-annotated corpus of Carlson et al. (2001) contains 238 CONTRAST relations that hold between two adjacent elementary discourse units. Of these, only 61 are marked by a cue phrase, which means that a program trained only on Carlson et al.'s corpus could identify at most 61/238 of the CONTRAST relations correctly. Because Carlson et al.'s corpus is small, all unmarked relations will be likely labeled as ELABORATIONs. However, when we run our CONTRAST vs. ELABORATION classifier on these examples, we can label correctly 60 of the 61 cue-phrase marked relations and, in addition, we can also label 123 of the 177 relations that are not marked explicitly with cue phrases. This means that our classifier contributes to an increase in accuracy from $61/238 = 26\%$ to $(60 + 123)/238 = 77\%$!!! Similarly, out of the 307 CAUSE-EXPLANATION-EVIDENCE relations that hold between two discourse units in Carlson et al.'s corpus, only 79 are explicitly marked. A program trained only on Carlson et al.'s corpus, would, therefore, identify at most 79 of the 307 relations correctly. When we run our CAUSE-EXPLANATION-EVIDENCE vs. ELABORATION classifier on these examples, we labeled correctly 73 of the 79 cue-phrase-marked relations and 102 of

the 228 unmarked relations. This corresponds to an increase in accuracy from $79/307 = 26\%$ to $(73 + 102)/307 = 57\%$.

5 Discussion

In a seminal paper, Banko and Brill (2001) have recently shown that massive amounts of data can be used to significantly increase the performance of confusion set disambiguators. In our paper, we show that massive amounts of data can have a major impact on discourse processing research as well. Our experiments show that discourse relation classifiers that use very simple features achieve unexpectedly high levels of performance when trained on extremely large data sets. Developing lower-noise methods for automatically collecting training data and discovering features of higher predictive power for discourse relation classification than the features presented in this paper appear to be research avenues that are worthwhile to pursue.

Over the last thirty years, the nature, number, and taxonomy of discourse relations have been among the most controversial issues in text/discourse linguistics. This paper does not settle the controversy. Rather, it raises some new, interesting questions because the lexical patterns learned by our algorithms can be interpreted as empirical proof of existence for discourse relations. If text production was not governed by any rules above the sentence level, we should have not been able to improve on any of the baselines in our experiments. Our results suggest that it may be possible to develop fully automatic techniques for defining empirically justified discourse relations.

Acknowledgments. This work was supported by the National Science Foundation under grant number IIS-0097846 and by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number MDA908-02-C-0007.

References

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France, July 6–11.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Aalborg, Denmark.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics NAACL–2000*, pages 132–139, Seattle, Washington, April 29 – May 3.

DUC–2002. *Proceedings of the Second Document Understanding Conference*, Philadelphia, PA, July.

Christiane Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. The MIT Press.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Jerry R. Hobbs. 1990. *Literature and Cognition*. CSLI Lecture Notes Number 21.

Eduard H. Hovy and Elisabeth Maier. 1993. Parsimonious or profligate: How many and which discourse structure relations? Unpublished Manuscript.

Alistair Knott and Ted J.M. Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30:135–175.

Alistair Knott, Jon Oberlander, Mick O'Donnell, and Chris Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text representation: linguistic and psycholinguistic aspects*, pages 181–196. Benjamins.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations, and common sense entailment. *Linguistics and Philosophy*, 16(5):437–493.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

James R. Martin. 1992. *English Text. System and Structure*. John Benjamin Publishing Company.

Deborah Schiffrin. 1987. *Discourse Markers*. Cambridge University Press.

TREC–2001. *Proceedings of the Text Retrieval Conference*, November. The Question-Answering Track.